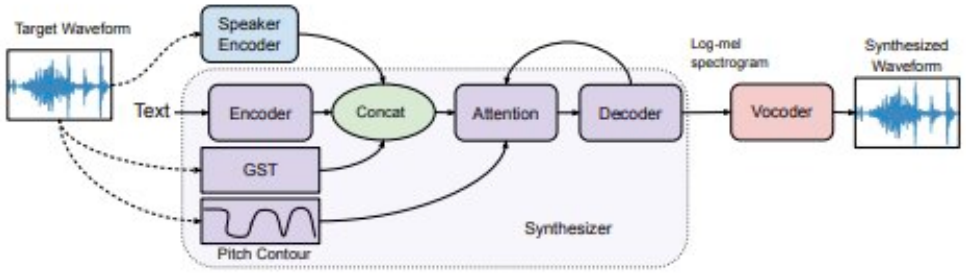
New method to make AI-generated voices more expressive

January 5 2022



Credit: University of California - San Diego

Researchers have found a way to make AI-generated voices, such as digital personal assistants, more expressive, with a minimum amount of training. The method, which translates text to speech, can also be applied to voices that were never part of the system's training set.

The team of computer scientists and electrical engineers from the University of California San Diego presented their work at the ACML 2021 conference, which took place online recently.

In addition to personal assistants for smartphones, homes and cars, the method could help improve voice-overs in animated movies, automatic translation of speech in multiple languages—and more.The method

could also help create personalized speech interfaces that empower individuals who have lost the ability to speak, similar to the computerized voice that Stephen Hawking used to communicate, but far more expressive.

"We have been working in this area for a fairly long period of time," said Shehzeen Hussain, a Ph.D. student at the UC San Diego Jacobs School of Engineering and one of the paper's lead authors. "We wanted to look at the challenge of not just synthesizing speech but of adding expressive meaning to that speech."

Existing methods fall short of this work in two ways. Some systems can synthesize expressive speech for a specific speaker by using several hours of training data for that speaker. Others can synthesize speech from only a few minutes of speech data from a speaker never encountered before; but they are not able to generate expressive speech and only translate text to speech. By contrast, method developed by the UC San Diego team is the only one that can generate with minimal training expressive speech for a subject that has not been part of its training set.

The researchers flagged the pitch and rhythm of the speech in training samples, as a proxy for emotion. This allowed their cloning system to generate expressive speech with minimal training, even for voices it had never encountered before.

"We demonstrate that our proposed model can make a new voice express, emote, sing or copy the style of a given reference speech," the researchers write.

Their method can learn speech directly from text; reconstruct a speech sample from a target speaker; and transfer the pitch and rhythm of speech from a different expressive speaker into cloned speech for the

target speaker.

The team is aware that their work could be used to make deepfake videos and audio clips more accurate and persuasive. As a result, they plan to release their code with a watermark that will identify the speech created by their method as cloned.

"Expressive voice cloning would become a threat if you could make natural intonations," said Paarth Neekhara, the paper's other lead author and a Ph.D. student in computer science at the Jacobs School. "The more important challenge to address is detection of these media and we will be focusing on that next."

The method itself still needs to be improved. It is biased toward English speakers and struggles with speakers with a strong accent.

More information: Paarth Neekhara et al, Expressive Neural Voice Cloning. arXiv:2102.00151v1 [cs.SD], arxiv.org/abs/2102.00151

Audio examples: expressivecloning.github.io/

Provided by University of California - San Diego